Some Useful Statistical Methods for Model Validation

Allan H. Marcus and Robert W. Elias

National Center for Environmental Assessment, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina

Although formal hypothesis tests provide a convenient framework for displaying the statistical results of empirical comparisons, standard tests should not be used without consideration of underlying measurement error structure. As part of the validation process, predictions of individual blood lead concentrations from models with site-specific input parameters are often compared with blood lead concentrations measured in field studies that also report lead concentrations in environmental media (soil, dust, water, paint) as surrogates for exposure. Measurements of these environmental media are subject to several sources of variability, including temporal and spatial sampling, sample preparation and chemical analysis, and data entry or recording. Adjustments for measurement error must be made before statistical tests can be used to empirically compare environmental data with model predictions. This report illustrates the effect of measurement error correction using a real dataset of child blood lead concentrations for an undisclosed midwestern community. We illustrate both the apparent failure of some standard regression tests and the success of adjustment of such tests for measurement error using the SIMEX (simulation–extrapolation) procedure. This procedure adds simulated measurement error to model predictions and then subtracts the total measurement error, analogous to the method of standard additions used by analytical chemists. — *Environ Health Perspect* 106(Suppl 6): 1541–1550 (1998). *http://ehpnet1.niehs.nih.gov/docs/1998/Suppl-6/1541-1550marcus/abstract.html*

Key words: model validation, measurement error

This paper discusses the application of a statistical procedure—measurement error correction—that allows essential adjustments to empirical comparisons between observed and predicted data. There is little question that some empirical comparisons are needed to validate a model (i.e., to assess the adequacy of the model to predict the outcome of proposed interventions). Measurement error correction removes some of the biases associated with these empirical comparisons. As part of the process of validation, this statistical procedure strengthens the confidence in the model.

For the integrated exposure uptake biokinetic (IEUBK) model, the environmental exposure media are air, water, diet, residential yard soil, and residential dust including multiple source contributions from paint, school or day care, and secondary residences. Measurements of exposure to any of these media are likely to be inaccurate estimates of actual exposure because of such factors as analytical error, repeat sampling variability, and location variability, and are not likely to completely characterize a child's actual long-term lead intake from that medium at that particular point in time. This measurement error is likely to be large enough to substantially attenuate the estimated relationship between observed blood lead and blood lead that is predicted from the model using the noisy input variables associated with that child's exposure. Similar effects are likely to occur in all modeling efforts, including the linear slope factor models that have been developed for long-term adult lead exposure by Bowers et al. (*1*) and by the U.S. Environmental Protection Agency (U.S. EPA) (*2*). In the worst case, measurement error may completely obscure the relationship between observed and predicted blood lead. Off-the-shelf statistical remedies for the problem of measurement error correction are not readily available. For the simple regression comparisons, the simulation and extrapolation (SIMEX) method proposed by Carroll et al. (*3*) may be adequate to estimate the true parameters relating observed and predicted values.

When empirical data are used to evaluate the model, there are several conventional statistical tests that can be applied to test the null hypothesis that the model output is wrong. These involve showing that some form of the predicted value does not equal the same form of the observed value (e.g., typical predicted ≠ typical observed). The simplest empirical comparison is that of a regression of observed values on predicted values (the usual variables are blood lead concentration, or the logarithm of blood lead, or the exceedance of blood lead over a health-based level of concern). If the usual assumptions of normal residuals and linearity are satisfied, we would test slope = 1, intercept = 0. But even in a well-calibrated model, when evaluated against an independent dataset, the typical result is slope less than 1, intercept greater than 0, even when the observed and predicted means are equal. The most plausible explanation, in our opinion, is that the data that are generally available as input for such models are not concurrent measurements of lead concentrations or loadings in environmental media in the residential or occupational setting in which that individual subject is believed to receive the exposure measured as blood lead.

If the purpose of the regression comparison is a formal test of the hypothesis (slope = 1, mean [observed] = mean [predicted]), then the distributional properties (normal, log-normal, etc.) of the adjusted estimates samples are critical in making accurate inferences. This may be even more critical in the logistic regression version of the test, comparing predicted and observed incidence of elevated blood leads, which also requires additional model assumptions about the intrinsic inter- and intraindividual variability of blood lead.

Measurement errors include sampling and analytical biases, instrument reading and recording errors, and temporal and spatial sampling sample collection discrepancies.

This paper is based on a presentation at the Workshop on Model Validation Concepts and Their Application to Lead Models held 21–23 October 1996 in Chapel Hill, North Carolina. Manuscript received at *EHP* 18 February 1998; accepted 6 July 1998.

Address correspondence to A.H. Marcus, National Center for Environmental Assessment (MD-52), U.S. Environmental Protection Agency, Research Triangle Park, NC 27711. Telephone: (919) 541-0636. Fax: (919) 541-1818. E-mail: marcus.allan@epa.gov

Abbreviations used: FNR, false negative rate; FPR, false positive rate; GSD, geometric standard deviation; IEUBK model, integrated exposure uptake biokinetic model for lead in children; LOC, a risk assessment term for level of concern; OLS, ordinary least squares, a statistical method for fitting a regression line; SIMEX, simulation and extrapolation, a statistical method that estimates measurement error and adjusts models to compensate for the measurement error; U.S. EPA, U.S. Environmental Protection Agency.

These errors may show diverse forms with diverse consequences, but in general are likely to introduce distortions of the predicted values without regard to the form of the predictive model. On the other side, the outcome measures against which the predictions are to be compared, in this case blood lead concentration, are also subject to measurement errors, including sampling and subject selection biases.

Limitations of Statistical Hypothesis Tests

Because of our concerns about the validity of formal statistical tests in the face of measurement error of unknown attributes, the validation strategy document for the IEUBK model (*4*) recommends that formal pass–fail statistical tests not be applied in empirical comparisons. The role of hypothesis testing in scientific inference has been hotly debated since its earliest uses, and remains a controversial subject for both statisticians and the subject-matter scientists who use statistical methods. However, if caution is used, formal hypothesis-testing methods for predictive models may be extremely helpful in diagnostic studies that estimate the range of conditions beyond which one might encounter some model inadequacies, or circumstances in which supplementary information needs to be collected. We have elaborated on five major areas of concern.

- Observational data may not have been collected for the purposes of validating the model. With the notable exception of the lead isotope study in five adult males carried out by Rabinowitz et al. (*5–8*), few observational studies of child lead exposure at lead-contaminated sites have been carried out for the purpose of validating any specific parameter or group of parameters in a predictive model. Descriptive analyses of cross-sectional epidemiology studies have been performed, but data were not collected to validate model parameters. The models were fitted to cross-sectional epidemiologic studies with standard statistical curve-fitting approaches, but the parameters estimated from the models did not usually have any comparable biologic counterpart. Representativeness, generalizability, and sample protocols are other serious issues that restrict the use of these data for model evaluation.
- The sample size may be too small, allowing inadequate power to detect model deficiencies or to discriminate among competing models. If the number of complete cases (paired blood lead and environmental lead data sufficient for model fitting and evaluation) is too small, then the uncertainty about goodness-of-fit statistics (as expressed by standard errors or confidence intervals) will be very large. Even models that fit poorly will not be distinguishable from the null hypotheses that the model fit is adequate. The statistical tests will have little power to detect failures of the prediction model.
- The sample size may be so large that even useful, moderately predictive models may be rejected by a statistical test for deviations that have little practical importance. If the sample size is very large, then even relatively small deviations of the model values from observed values are likely to be declared statistically significant. Suppose, for example, that no water lead samples are collected in an epidemiology study. The IEUBK model may assume a standard lead intake of about 0.5 liters per day at a concentration of 4 µg/liter, or about 2 µg/day. If in fact the tap water lead concentration is negligible, then the excess of slightly less than 2 µg/day in the IEUBK model predictions would probably be significant in studies with more than 500 to 1000 children.
- Measurement of exposure may be inaccurate, biasing many standard statistical tests. The potentially serious difficulties that exposure measurement errors may cause has only recently been recognized. Measurement errors in exposure variables and other covariates used as model input can propagate through the model and produce an inaccurate model prediction. Carroll and Galindo (*9*) illustrate conclusively how measurement error can distort the apparent relation between exposure and biologic response, and will very probably bias the test statistics in the direction of attenuating the apparent predictiveness of the model.
- Blood lead is not necessarily a "gold standard" for model evaluation. Temporal and behavioral influences on exposure, such as the season of the year, the level of information that the child, the child's parents, or caretakers have about lead hazards, and the amount of time that the child spends away from home can be significant modifiers of exposure in determining child blood lead. Seasonal rhythms of blood lead concentrations discussed by the U.S. EPA (*10*), Hogan et al. (*11*), and Mushak (*12*) in this monograph provide many examples of the assessment and interpretation of several modifying factors in evaluation of lead models.

Classical statistical tests mistakenly assume that the predicted values, prediction intervals, and classification of elevated blood lead concentrations based on the model are statistically accurate predictors of what they purport to predict. Both systematic and random errors in epidemiologic studies influence the accuracy of the predictors. Random errors may occur when single, individual samples do not take into account temporal variability; when spatial samples of yard soil and house dust do not represent the actual play areas and contact surfaces of the child; when exposure has been modified by environmental factors such as groundcover or dust loading; or by behavioral factors such as housecleaning practices, choice of play area, hand-washing frequency, and mouthing of nonfood objects. Random errors of sample collection and processing may also occur when the sample is contaminated by other environmental media, when the sample is modified during transit and storage, or when the sample data are misrecorded.

Systematic errors might occur if instruments are miscalibrated, measurements are taken at inappropriate locations or seasons, no measurements or estimates are made of nonresidential exposure, or the subjects are not representative of the same sociodemographic or ethnic groups as the model.

Statistical Tests of Hypotheses That Evaluate Predicted Values

We will illustrate a few of the potential problems in applying formal statistical goodness-of-fit tests without considering the possible effects of measurement error. The simplest form is a linear regression test that asks the question: Does the observed value equal model-predicted value? Most formal evaluations ask some variant of this question. Frequently, a new variable is constructed to convert the hypothesis to the univariate form. Typical forms of the goodness-of-fit test are the following:

Does the observed value minus model-predicted value equal zero (showing that the predictions are unbiased)?

Does the ratio of observed value to model-predicted value equal one (in which case the predictions are relatively unbiased)?

Does the log of the observed value minus the log of the model-predicted value equal zero (in which case the predictions are relatively unbiased)?

In order to establish a general notation for these tests, the algebraic notation is used here:

$$d = Y - M, \qquad [1]$$

where

d = prediction error,

Y = observed value (e.g., blood lead or log blood lead), and

M = modeled value analogous to Y (M is a model prediction not necessarily derived from an optimized fit of observed values).

Of the many possible statistical hypotheses of model adequacy that can be tested using a set of paired values of an observation Y and its model prediction M, five are listed below as null hypotheses $H_0(1)$ to $H_0(5)$. Figure 1 illustrates the graphical interpretation of these five statistical hypotheses, plus two additional hypotheses described in the next section.

$H_0(1)$: mean $d = 0$

Hypothesis $H_0(1)$ expresses a common concept: Although some differences between observed values and predicted values are expected, there should be no difference between the mean observation and the mean prediction from a good model. Mean could be replaced by some other measure of typical value, such as the median or geometric mean, if prediction errors (d) have an asymmetric or heavy-tailed distribution.

$H_0(2)$: mean Y = mean M

Or, if Y = log blood lead, then $H_0(2)$: geometric mean Y = geometric mean M.

Hypothesis $H_0(2)$ is more general than $H_0(1)$ and addresses a common situation in epidemiologic studies in which data sets may have some records in which the data are not paired. That is, the environmental measurements for calculating a value of M are available, but not a corresponding blood lead observation Y; or conversely, Y is available, but there is not enough environmental data to calculate a corresponding value of M. If the missing values are missing completely at random and the existing data are representative of the missing data, then the observed and modeled

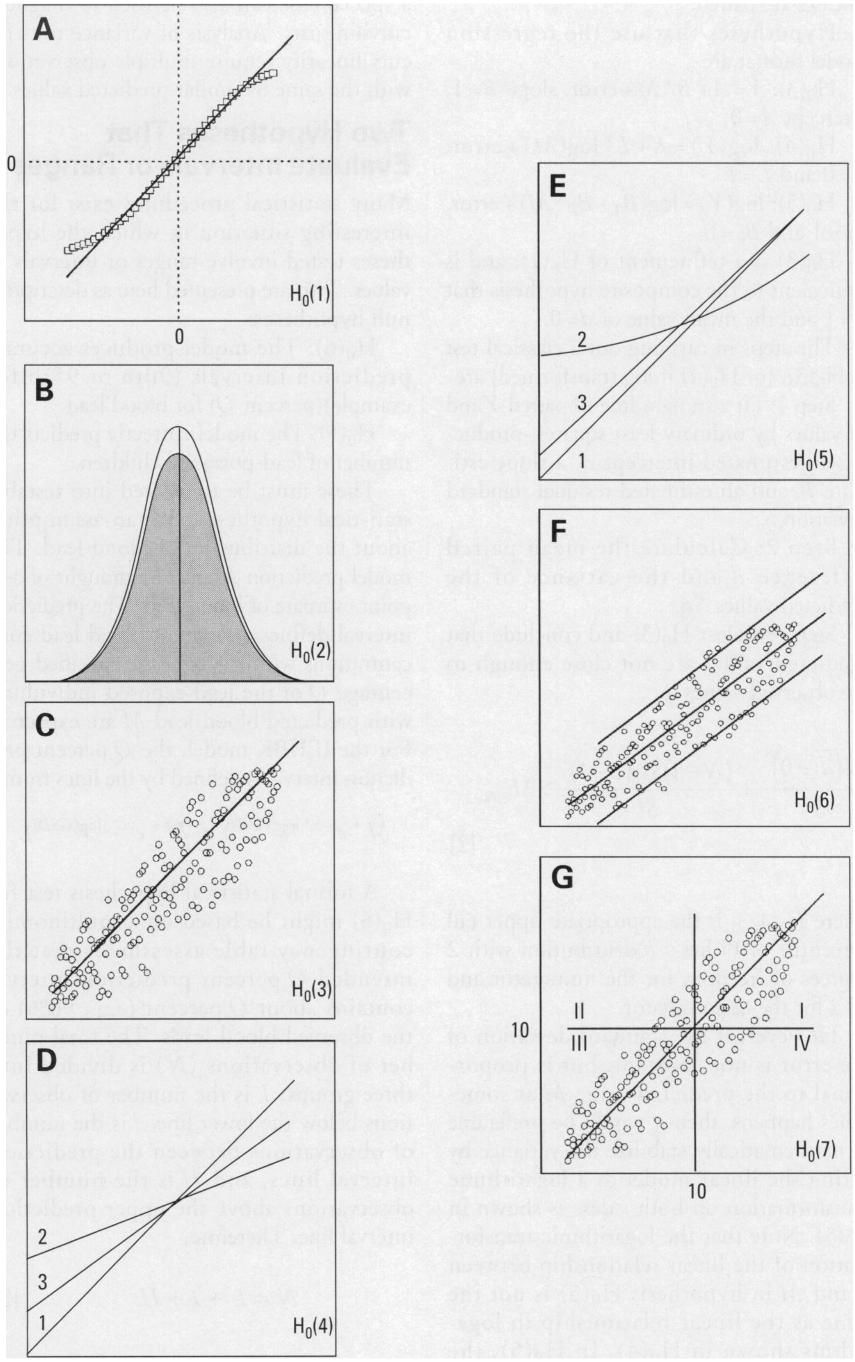

Figure 1. Diagrammatic illustration of the statistical tests for hypotheses $H_0(1)$ through $H_0(7)$. (*A*) illustrates the condition in $H_0(1)$ where the difference between observed and predicted blood lead concentrations is expected to equal zero. (*B*) illustrates $H_0(2)$; the observed mean equals the predicted mean. (*C*) expands $H_0(1)$ to show that, according to $H_0(3)$, the slope of observed vs. predicted should equal one and pass through the origin. (*D,E*) represent $H_0(4)$ and $H_0(5)$, which are two different logarithmic transformations of $H_0(3)$, and test the condition that the slope = 1 and intercept = 0. These are the two tests that are used in this paper for the SIMEX procedure for measurement error adjustments. (*F,G*) illustrate two additional hypotheses that evaluate intervals and ranges, both of which are important for risk assessment. Hypothesis $H_0(6)$ tests the accuracy of predicting a specific upper tail interval (e.g., 95%), and hypothesis $H_0(7)$ tests the validity of the prediction of the number of children with blood lead concentrations exceeding 10 µg/dl. Correctly predicted blood lead concentrations are in quadrants I and III, incorrectly predicted are in II and IV.

datasets can be compared even though not all cases are paired.

Hypotheses that use the regression model format are

$H_0(3)$: $Y = A + B^* M +$ error, slope $B = 1$, intercept $A = 0$

$H_0(4)$: $\log(Y) = K + L^* \log(M) +$ error, $K = 0$ and $L = 1$

$H_0(5)$: $\log(Y) = \log(B_0 + B_P{}^* M) +$ error, $B_P = 1$ and $B_0 = 0$

$H_0(3)$ is a refinement of $H_0(1)$, and is equivalent to the composite hypothesis that $B = 1$ and the mean value of $d = 0$.

The steps in carrying out a classical test of $H_0(3)$, [or $H_0(4)$ if log transformed] are

Step 1: Fit a straight line to paired Y and M values by ordinary least squares, producing an estimated intercept A, a slope estimate B, and an estimated residual standard deviation S.

Step 2: Calculate the mean paired difference $\bar{d}$ and the variance of the predicted values S_M^2.

Step 3: Reject $H_0(3)$ and conclude that predicted values are not close enough to the observed values if

$$\frac{N(\bar{d}-0)^2}{S^2} + \frac{(N-1)S_M^2(B-1)^2}{S^2} > 2F_{2,N-2} \qquad [2]$$

where $F_{(2,\,N-2)}$ is the appropriate upper tail percentile of Fisher's F distribution with 2 degrees of freedom for the numerator and N–2 for the denominator.

However, if the standard deviation of the error is not constant, but is proportional to the predicted value M as sometimes happens, then it would be preferable to mathematically stabilize the variance by fitting the linear model to a logarithmic transformation on both sides, as shown in $H_0(5)$. Note that the logarithmic transformation of the linear relationship between Y and M in hypothesis $H_0(5)$ is not the same as the linear relationship in logarithms shown in $H_0(4)$. In $H_0(5)$, the hypothesis is that the relationship between Y and M is linear at all values of M but will not pass through the origin ($Y = 0$ when $M = 0$) unless $A = 0$. In $H_0(4)$, the hypothesis is that the relationship between Y and M always passes through the origin but is linear only if $L = 1$. In this respect, all three hypotheses, $H_0(3)$ through $H_0(5)$, test different sets of assumptions.

As with any regression analysis, residuals should be carefully examined for outliers, curvilinearity, and trends. Nonlinear parametric models that include linearity as a special case can also be used to diagnose curvilinearity. Analysis of variance tests for curvilinearity require multiple observations with the same or similar predicted values.

Two Hypotheses That Evaluate Intervals or Ranges

Many statistical procedures exist for the interesting situation in which the hypotheses tested involve ranges or intervals of values. Two are presented here as descriptive null hypotheses:

$H_0(6)$: The model produces accurate prediction intervals (90th or 95th for example)(percent Q) for blood lead.

$H_0(7)$: The model correctly predicts the number of lead-poisoned children.

These must be translated into testable statistical hypotheses with an assumption about the distribution of blood lead. The model prediction M may be thought of as a point estimate of blood lead. The prediction interval defines a range of blood lead concentrations within which the specified percentage Q of the lead-exposed individuals with predicted blood lead M are expected. For the IEUBK model, the Q percent prediction interval is defined by the lines from

$$M * e^{(-z * \log(GSD))} \text{ to } M * e^{(z * (\log(GSD))}.$$

A formal statistical hypothesis test for $H_0(6)$ might be based on a multinomial contingency table assessment that the intended Q percent prediction interval contains about Q percent (e.g., 90%) of the observed blood leads. The total number of observations (N) is divided into three groups: L is the number of observations below the lower line, I is the number of observations between the prediction interval lines, and H is the number of observations above the upper prediction interval line. Therefore,

$$N = L + I + H. \qquad [3]$$

The null hypothesis would then have the form

$$E\{L\} = E\{H\} = N\left(\frac{(100 - Q)}{200}\right). \qquad [4]$$

The statistical translation of $H_0(7)$ is more difficult. A useful tabular framework is shown in Table 1. The blood lead level of concern (LOC) is defined by criteria described by the Centers for Disease Control and Prevention (*13*). Elevated blood lead means any blood lead concentration that is at least as large as the LOC. Table 1 uses the following definitions:

A = number of children with observed and predicted blood leads is less than the LOC;

B = number of children with elevated blood lead and predicted blood lead less than the LOC;

C = number of children with blood lead less than the LOC predicted to have blood lead equal to or less than the LOC;

D = number of children with both observed and predicted elevated blood lead equal to or less than the LOC.

Many appropriate figures of merit can be calculated from this table. A and D are accurate classifications that we wish to maximize, B and C are inaccurate classifications that we wish to minimize. In classical epidemiology terms (*14*), sensitivity is the proportion of children with elevated blood lead that will be classified correctly by the prediction model (Equation 5), and specificity is the proportion of children without elevated blood lead who are correctly classified by the prediction model (Equation 6). Many investigators are also concerned about the false positive rate (proportion of children classified as likely to have elevated blood lead who are observed to have nonelevated blood lead), denoted FPR, and the false negative rate (proportion of children classified as likely to have nonelevated blood lead who are observed to have elevated blood lead), denoted FNR. These can be calculated from Table 1 as

$$\text{Specificity} = \frac{A}{(A+C)} = \frac{A}{U} \qquad [5]$$

$$\text{Sensitivity} = \frac{D}{(B+D)} = \frac{D}{V} \qquad [6]$$

$$\text{FNR} = \frac{B}{(A+B)} = \frac{B}{S} \qquad [7]$$

$$\text{FPR} = \frac{C}{(C+D)} = \frac{C}{T} \qquad [8]$$

There is clearly a trade-off among these criteria, which can be optimized by combining them into a single index or criterion based on, for example, the costs of incorrect decisions (B or C) versus correct decisions (A or D). It is likely that many public health investigators would prefer to

Table 1. A 2 × 2 table for empirical comparisons of elevated blood lead.

	Observed < LOC	Observed > LOC	Row sum
Predicted < LOC	A	B	S
Predicted ≥ LOC	C	D	T
Column sum	U	V	*N* = total

minimize the FNR. The most obvious value-free requirement is that the table be symmetric, i.e., $E\{B\} = E\{C\}$, where $E\{\ \}$ is the expected value of the variable (*15*). Formulating these hypotheses suggests useful ways to display data for empirical comparisons. The next section will demonstrate why we recommend that great care should be used when actually performing any of these tests.

Our concerns are not merely hypothetical. Neither the blood lead data used for comparisons nor the input data used in model predictions can be assumed to be without blemish. Errors in data used as model input can seriously distort formal statistical tests that may be used in model evaluation. Carroll et al. (*3*) proposed a much more detailed discussion of these effects. Summarized briefly, noisy input data can distort the empirical comparison in either direction but usually in the direction of attenuating the apparent predictiveness of the model. They describe a statistical methodology for removing some of the distortion. Stepping through the problem systematically, we note that differences between observed and predicted values generally have greater variability than variability in the observed values alone, due to input errors in the predicted values. The model propagates this uncertainty about input values into uncertainty about the model output. The effects may be characterized mathematically (see Equation 1):

d = observed value–noisy modeled value, which can be expanded to

d = (observed value – true modeled value) + (true modeled value – noisy modeled value).

In general, the second term may be expected to add both random and systematic biases to an empirical comparison structured like $H_0(1)$ or $H_0(2)$. The consequences are more serious for regression-structured evaluations such as $H_0(3)$ through $H_0(5)$. Several authors (*3,9,16,*) demonstrated that the simple regression of observed on predicted value with noisy model predictions caused by input errors will attenuate the slope (*B* or *L*) of the linear regression of the observed values *Y* on the corresponding predictions *M*. Both the regression and the correlation coefficients assume values closer to zero than the true values, and there would be a corresponding change in the intercept terms (*A* or *K*). The usual case is that the slope estimate *B* or *L* decreases to a value less than 1, with a corresponding intercept *A* or *K* greater than zero. This may imply that model does not adequately predict the observations. Furthermore, regression tests on data with larger variability and larger standard errors for parameters may produce lower significance in hypothesis tests. Tests of a linear versus nonlinear relationship between *Y* and *M* may also be distorted, usually toward a more linear relationship than really exists.

Similar effects occur when regression comparisons are made of logistic (binary) and categorical (grouped) data. Such tests are usually performed when numeric differences of observed and predicted values are replaced by indicator variables such as coding 1 for inside and 0 for outside the prediction intervals in testing $H_0(6)$. Likewise, $H_0(7)$ might be evaluated by coding observed blood lead less than the LOC as 0 and elevated blood lead as 1, and regressing these on predicted risk or logits for elevated blood lead for each subject. Predictor measurement error will distort these comparisons.

Finally, even the contingency table formulations of $H_0(6)$ or $H_0(7)$ shown in Table 2 are likely to be biased because the predicted values will be misclassified into the wrong category.

A Numerical Example: IEUBK Model Comparisons

We use the dataset that was evaluated by Hogan et al. (*11*) to focus the reader's attention on measurement error correction and other theoretical aspects of the methodological issues of comparing model predictions with observed blood lead data, not on a particular model or a particular epidemiology study. The dataset contains a large number of observations from a cross-sectional epidemiology study, with particular emphasis on children less than 6 years of age. We demonstrate several tests of the IEUBK model. The tests would be equally appropriate for any other predictive child blood lead model with similar input data.

The IEUBK model is intended to describe the distribution of blood lead concentrations expected when all sources of the child's environmental lead exposure have been identified. The data, however, only contain information about the child's residential lead exposure. Therefore, for the purposes of demonstrating some of the statistical evaluation methods described in the preceding section, we use some ancillary information (i.e., the number of hours per week that the caretaker reported the child as present at home). The majority of cases were reported to spend all of the time (168 hr/week) at home. It is highly unlikely that all these children spent all their time inside or in the immediate vicinity of their residence. On the other hand, it is likely that for most of the time these children spent in other locations, the lead exposure was essentially the same. Therefore, we report only the records for the 282 children who met these criteria, and which had sufficient data (age, soil lead or house dust lead, blood lead) to allow calculation of an IEUBK-predicted blood lead, and empirical comparison with observed blood lead.

Preliminary Evaluation and Data Screening

The observed logarithms of blood lead are shown in Figure 2 against the IEUBK predictions, with 80% prediction intervals derived from the IEUBK model run, assuming the geometric standard deviation (GSD) of 1.6. The line log(observed) = log(predicted) is shown at the center of the

Table 2. Observed versus predicted blood lead by level-of-concern category.

	Observed blood lead				
	< 10	10–14	15–19	20+	Row sum
Predicted blood lead					
< 10	184	24	9	6	223
10–14	25	4	2	2	33
15–19	8	1	0	2	11
20+	5	2	0	1	8
Column sum	222	31	11	11	275

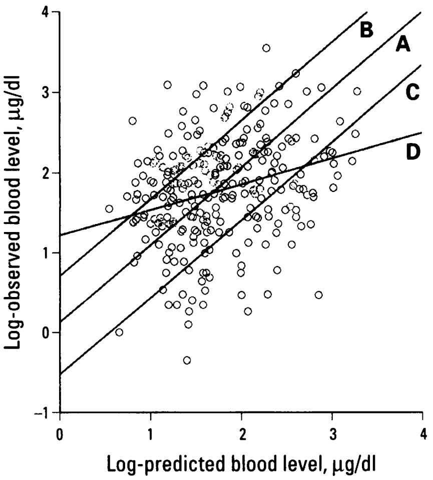

Figure 2. Comparison of observed log(blood lead) vs. IEUBK model-predicted log(blood lead) in 275 children who were reported as spending all their time at home. The upper and lower parallel lines (B,C) represent the IEUBK 80% prediction interval; the middle parallel line (A) is the null hypothesis (observed = predicted). Line D is the OLS regression line.

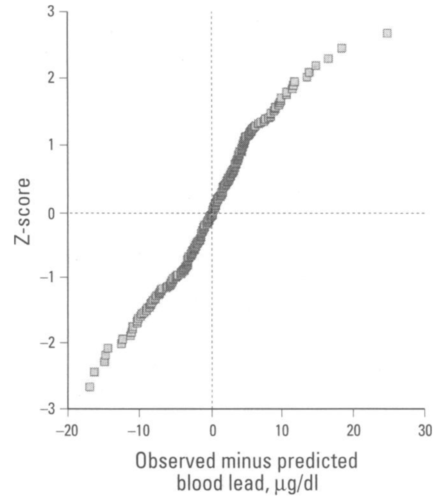

Figure 3. Cumulative probability plot of the difference, observed blood lead minus IEUBK model-predicted blood lead in 275 children who were reported as spending all their time at home. Vertical axis is on a normal or Gaussian probability scale (Z score). The range from $Z = \pm 1$ includes the central 68% of distributions, and the range from $Z = \pm 1.96$ includes central 95% of distribution.

interval, corresponding to $H_0(5)$ with $K = 0$ and $L = 1$. Figure 2 shows only 275 points. Based on several tests, we deleted seven points that appear to be outliers.

Differences between Observed and Predicted Values

The normal probability plot of the cumulative distribution is shown in Figure 3. The central 68 to 70% (from $z = -1$ to $z = 1$) is nearly linear. The upper and lower tails, however, are linear with a much flatter slope. This suggests that the differences are not normally distributed, but might be the mixture of at least two roughly normal distributions, one with much greater variability than the other.

Empirical Comparisons Using Counting Data

Tables 2 through 4 show several other comparisons that may be useful alternatives in presenting the results. Table 2 indicates the extent to which the predictions are, on the whole, unbiased: the number of predicted values higher than the observed in any given blood lead category is about the same as the number of observed values higher than the predicted values in the analogous (transposed) category. Table 3 reduces the information in Table 2 into three 2×2 tables, again showing the desired symmetry or lack of significant bias.

Table 4 shows how the graphical information in Figure 2 can be used in a formal test for the adequacy of a prediction interval. The visual impression from Figure 2 is that more than 20% of the observations lie outside the prediction interval. In fact, as shown in Table 4, only about 55% of the observations lie inside the 80% prediction intervals. This suggests, again, that there may be a subpopulation of children whose blood lead concentrations do not fit a lognormal distribution with a GSD of 1.6.

Adjusting the Regression Test for Measurement Error

The initial regression results deviate substantially from the null hypotheses $H_0(3)$ and $H_0(4)$, with a Figure 1 ordinary least-squares (OLS) slope of about 0.3 and an intercept of about 1.2 on the log scale (3.3 µg/dl). Clearly the hypothesis that the regression model of $Y = A + B^* M + \text{error}$ [$H_0(3)$] or its log form [$H_0(5)$] would be rejected if the slope deviates significantly from 1 and the intercept from 0. However, much of this deviation may be attributable to measurement error.

A useful and quite general method for dealing with measurement error in nonlinear regression has recently been proposed and shown to be generally valid (*3,17*). This is the SIMEX method. The concept is very simple: if measurement error biases the estimate, then adding more measurement error should increase the bias. The relationship between the expected value of the estimated coefficient (*B* or *L*) and the true coefficient, with and without measurement error, respectively, can be described well in large samples by the equation

$$E\{B\} = \frac{\beta}{\left(1+\frac{\sigma_M^2}{\sigma_P^2}\right)}, \qquad [9]$$

where $E\{B\}$ = expected value of the estimated coefficient

β = true coefficient

σ_M = measurement error standard deviation of the predictor

σ_P = standard deviation of the true predictor.

In this equation, as σ_M approaches zero, the expected value of the estimated coefficient approaches the true coefficient (β). The true predictors and the true predictor standard deviation cannot be observed because they depend on the true values of the input variables such as the true time-weighted and soil ingestion rate-weighted soil lead concentration, the true time-weighted and dust ingestion

Table 3. 2×2 tables of observed versus predicted elevated blood lead.

LOC = 10			LOC = 15			LOC = 20		
Predicted	Observed		Predicted	Observed		Predicted	Observed	
	<10	10+		<15	15+		<20	20+
<10	184	38	<15	237	16	<20	257	7
10+	39	14	15+	19	3	20+	10	1

Table 4. Comparison of observed blood lead with prediction intervals: 80% prediction intervals with GSD = 1.6.

Number	Below interval	Inside interval	Above interval	Total number
Observed	63	145	67	275
Expected[a]	27.5	220	27.5	275

[a] Expected numbers are calculated from 0.1 *N* below, 0.8 *N* inside, and 0.1 *N* observations above the prediction limits.

rate-weighted dust lead concentration, and so on, which cannot be truly measured. However, if some estimate of the model standard deviation, σ_P, is available, then the observed slopes such as B or L can be adjusted empirically by fitting the slope attenuation model to a set of simulated measurements that are even more noisy than the real data, and extrapolating the observations backward to the known or inferred value of σ_M, assigned a negative effect as shown below.

We demonstrate this method as a test of the linear empirical comparison regression model, fitted in a logarithmic form, shown above as $H_0(5)$. The OLS fit to predicted values would look like a straight line on nontransformed plot but as a curved line on a log–log scale. The SIMEX procedure was carried out by the following steps:

Step 1: Estimate the slope B_P and intercept B_0 in a nonlinear least-squares regression model log(observed blood lead) = log(B_0 + B_P* predicted blood lead) + error. We used SAS PROC NLIN (*18*).

Step 2: For each predicted value M, generate a standard normal random variate Z, and calculate a randomized predicted value with additional log-normally distributed error corresponding to M,

$$M_{ran} = M * e^{\left(Z\sigma_M\right)} \qquad [10]$$

This assumes that the measurement errors are log-normally distributed, with median or geometric mean equal to 1 and GSD = exp(σ_M). In these examples, we used σ_M in steps of 0.1 from 0.1 to 1.0. Note that σ_M is a purely hypothetical value that brackets the range of plausible measurement error in log(predicted blood lead) not the log GSD of the population of true measurement errors or the population of predicted values M.

Step 3: Simulation. Repeat Step 2 many times for each set of N simulated predictors M_{ran}. Figure 4 shows the 25 values of the slope B_P fitted by the $H_0(5)$ regression model to each pair of 275 data values of the observed blood lead versus simulated randomized predicted values with extra measurement error for each of 10 values of σ_M. Note that when the added error is small, then the estimated slope B_P is only slightly attenuated on average from the nonlinear *LS* slope at $\sigma_M = 0$. However, the slope estimate B_P is attenuated nearly to 0 at the upper end of the measurement error range. The range of σ_M values for the slope is large enough to sustain a good nonlinear regression model shown in Step 4.

Figure 5 shows the 25 values of the intercept B_0 fitted by the $H_0(5)$ regression model to each pair of 275 data values of the observed blood lead versus simulated randomized predicted values with extra measurement error, for each of 10 values of σ_M. Note that when the added error is small, then the estimated intercept B_0 is only slightly inflated on average from the nonlinear *LS* intercept of about 3.3 µg/dl at $\sigma_M = 0$. However, the intercept estimate B_0 is inflated nearly to the blood lead geometric mean of 5 µg/dl at the upper end of the measurement error range. The range of σ_M values for the intercept is also large

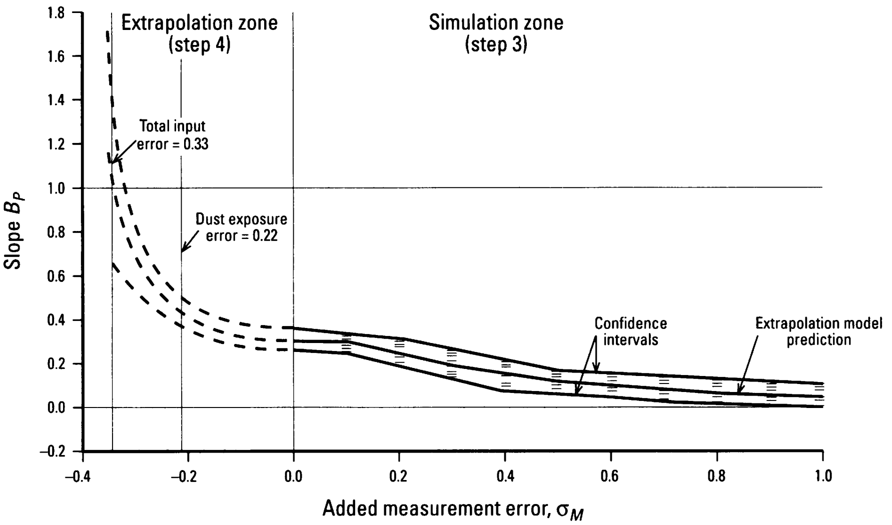

Figure 4. The slope B_P of the line log(observed) = log(B_0 + B_P* predicted) is shown versus the measurement error standard addition Σ for 250 simulated replicates of 275 cases. The nonlinear function fitted to B_P for $\Sigma \geq 0$ is shown in solid lines to the right of 0.0, along with interpolated 95% confidence intervals. The extrapolated values are shown in dashed lines to the left of 0.0 for $\Sigma < 0$. The intersections of the vertical lines with the upper and lower dashed curves show confidence intervals for B_P after adjustment for measurement errors attributable to dust alone [0.22 = log(1.25)] or to all environmental media [0.33 = log(1.39)].

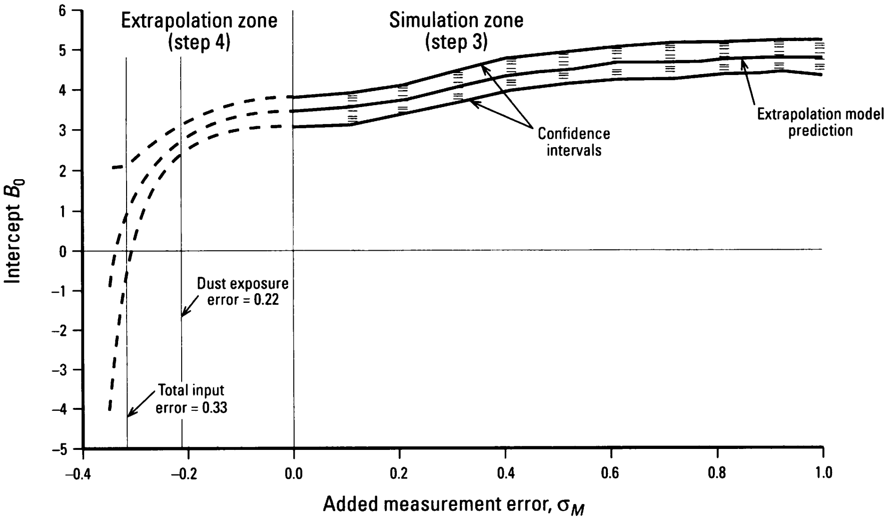

Figure 5. The intercept B_0 of the line log(observed) = log(B_0 + B_P* predicted) is shown vs. the measurement error standard addition "sigma" for 250 simulated replicates of 275 cases. The nonlinear function fitted to B_0 for $\Sigma \geq 0$ is shown in solid lines, along with interpolated 95% confidence intervals. The extrapolated values are shown in dashed lines for $\Sigma < 0$. The intersections of the vertical lines with the upper and lower dashed curves show confidence intervals for B_0 after adjustment for measurement errors attributable to dust alone [0.22 = log(1.25)] or to all environmental media [0.33 = log(1.39)].

enough to sustain a good nonlinear regression model shown in Step 4.

Step 4: Extrapolation. The extrapolation data set consists of 25 values of simulated B_0 and B_P pairs for each of 10 values of σ_M, and the nonlinear *LS* fit at $\sigma_M = 0$. We fitted the following nonlinear extrapolation models with parameters G_0, G_1, G_2 to the 251 $(25 \times 10 + 1)$ values of B_0:

$$B_0 = G_0 + \frac{G_1}{\left(G_2 + \sigma_M^2\right)} \qquad [11]$$

and an analogous model with parameters P_0, P_1, P_2 to the 251 values of B_P,

$$B_P = P_0 + \frac{P_1}{\left(P_2 + \sigma_M^2\right)} \qquad [12]$$

The values for the *G* and *P* parameters, which are outputs from the SAS PROC NLIN using the SIMEX method, are given in Table 5. The fitted models and their 95% confidence intervals are shown in Figures 4 and 5 as the smooth curves, covering the range of observed and simulated data. Several alternative fits were carried out, evaluating different weights and variance-stabilizing transformations. The models shown had weight 25 for the nonlinear *LS* value and no transformation, which produced the smallest extrapolation confidence bands.

The same models were then extrapolated by subtracting out hypothetical values of the true standard error. The extrapolation part of the analysis is shown by the dashed curves in Figures 4 and 5. These functions are given by:

$$\text{extrapolated } B_0 = G_0 + \frac{G_1}{\left(G_2 - \sigma_M^2\right)} \qquad [13]$$

$$\text{extrapolated } B_P = P_0 + \frac{P_1}{\left(P_2 - \sigma_M^2\right)} \qquad [14]$$

from the smooth-fitted curves. These functions would expand to $B_0 = -\infty$ and $B_P = \infty$ as σ_M^2 approaches G_2 or P_2. However, much smaller values of the true measurement error standard deviation σ_M are appropriate. The next step is to select the boundary conditions for σ_M^2 and apply these to the B_0 and B_P equations using the *G* and *P* parameters generated by the simulation.

The IEUBK model GSD value of 1.6 reflects a composite of measurement input errors, reflecting: *a*) environmental exposure concentration errors, and *b*) variability in biologic and behavioral factors reflected by different absorption coefficients, compartment volumes and transfer coefficients, intake and ingestion rates, and other idiosyncratic exposures. Variability in environmental exposure, denoted GSD_E, is the most appropriate component of measurement error to be evaluated for risk assessment. The environmental media concentrations or loadings are usually the most important health risk determination numbers that are factored into site-specific remediation decisions. Biologic and behavioral variability are unavoidable components, but we assume that these can be characterized together as a log-normal component with a GSD denoted GSD_B. Assuming that biologic and behavioral variability are independent of environmental measurement variability, then the following model applies:

$$GSD = e^{\sigma_M} \qquad [15]$$

$$GSD_B = e^{\sigma_B} \qquad [16]$$

$$GSD_E = e^{\sigma_E} \qquad [17]$$

$$\sigma_M^2 = \sigma_B^2 + \sigma_E^2 \qquad [18]$$

Reasonable ranges for GSD_B may be inferred from other environmental studies. Almost no human population or child subgroup study has identified a residual GSD smaller than about 1.3 to 1.4, which is a reasonable range for GSD_B. Furthermore, simulation studies on the propagation of measurement error through the IEUBK model (*19*) suggests that with reasonable uncertainties in dust lead alone, dust lead measurement error can induce a range of variability in predicted blood lead with a GSD_E of about 1.25 to 1.35, whereas with uncertainty in correlated dust lead and soil lead, a reasonable GSD_E is in the range 1.35 to 1.45. In Table 6, we show a small set of possible GSD_B values within the probable range, and the corresponding values for GSD_E . The total measurement error shown in Figures 4 and 5, $\sigma_M = \sigma_E = 0.33 = \log(1.39)$, is a reasonable choice. From Equation 13 the estimate of B_0 is not significantly different from 0, and from Equation 14, the estimate of B_P is not significantly different from 1, by conventional standards.

The relationship between observed blood lead and values predicted from the IEUBK model was much closer to the appropriate null hypothesis when the

Table 5. Extrapolation parameters (G and P) from nonlinear least-squares analysis.

Alternative hypothesis model	Hypothesis parameter	Extrapolation		Value at σ_M=0.33	Value at σ_M=0	Maximum σ_M
		Parameter	Estimate[a]			
H_0(4)						
log–log	L_0	G_0	1.712	0.50	1.05	0.40
		G_1	–0.158			
		G_2	0.239			
	L_P	P_0	0.001	0.70	0.38	0.41
		P_1	0.0885			
		P_2	0.236			
H_0(5)						
log of linear	B_0	G_0	5.85	0.21	3.47	0.44
		G_1	–0.472			
		G_2	0.198			
	B_P	P_0	–0.031	1.00	0.41	0.44
		P_1	0.0856			
		P_2	0.1922			

[a] Assuming lead in dust greater than 2000 ppm imputed at 2000 ppm.

Table 6. Sensitivity analyses for environmental measurement error.

GSD	σ = log (GSD)	GSD_B	σ_B = log (GSD_B)	σ_E[a]	GSD_E
1.6	0.4700	1.30	0.2624	0.3900	1.4770
1.6	0.4700	1.35	0.3001	0.3617	1.4357
1.6	0.4700	1.40	0.3365	0.3282	1.3884

[a] σ_E was calculated from equation 19: $\sigma_E = \sqrt{\sigma^2 - \sigma_B^2}$.

linear regression models were adjusted for measurement error. The two models [hypotheses $H_0(4)$ and $H_0(5)$] are compared in Figure 6. Hypothesis $H_0(4)$ assumes a linear regression for log(observed blood lead) versus log(model blood lead), and hypothesis $H_0(5)$ assumes a linear relationship between observed and modeled blood, which is fitted after logarithmic transformation of both sides.

On the log–log plot of Figure 6, the hypothesis $H_0(4)$ alternatives are straight lines. Note that when there is no adjustment for measurement error, the unadjusted OLS fit has an intercept of 0.9 and a slope of 0.4, whereas after adjustment for measurement error with log(GSD) = 0.33, the SIMEX adjustment gives an intercept of 0.5 and a slope of 0.7. This is much closer to null hypothesis, although the difference between observed and predicted blood lead is still substantial for some risk assessment applications. A measurement error GSD larger than 1.4 in the model values would be needed to bring the curves closer, and cannot be justified based on empirical evidence discussed earlier.

The hypothesis $H_0(5)$ alternatives are curved lines on Figure 6. Note that when there is no adjustment for measurement error, the OLS fit gives an intercept of 3 and a slope of 0.47, whereas after adjustment for measurement error with log(GSD) = 0.33, the SIMEX adjustment gives am intercept of 0.2 and a slope of 1.0. This is much closer to null hypothesis line (observed = predicted), with an expected intercept of 0 and slope of 1.0, and the difference between observed and predicted blood lead using this measurement error correction method is negligible for risk assessment applications. A measurement error GSD of about 1.4 in the model values seems to be appropriate. The IEUBK model is only slightly nonlinear at blood lead less than 25, so that the family of linear alternatives in hypothesis $H_0(5)$ may be more realistic. An important aspect of this procedure is that no individual observed values were changed. The variability due to measurement error was enhanced by a method similar to standard additions, then extrapolated to a preselected value for the GSD using an equation derived from a SIMEX application of SAS PROC NLIN.

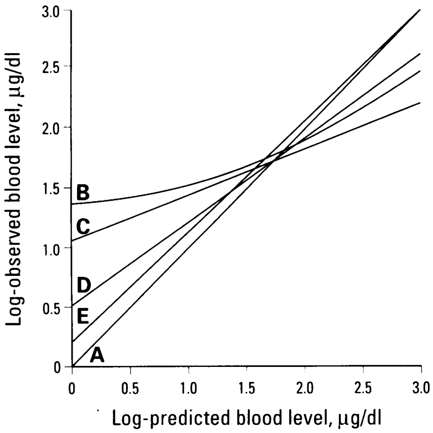

Figure 6. The application of the SIMEX measurement error correction procedure to two hypotheses, $H_0(4)$ and $H_0(5)$. Line A is the theoretical condition where observed = predicted (intercept = 0, slope = 1). Lines B and C are the uncorrected and corrected forms of $H_0(4)$, and lines D and E are the uncorrected and corrected forms of $H_0(5)$.

We may therefore accept the statistical hypothesis and conclude that with corrections for measurement error, the IEUBK model provides a satisfactory prediction of typical blood lead concentrations for children exposed to residential lead in this residential situation. This process also raises the possibility that there is a small subpopulation of children with blood lead concentrations either much higher or much lower than those predicted by the model with a standard GSD of 1.6.

Conclusions

Hypothesis tests can be a useful statistical tool for model validation. Several forms of statistical hypotheses were presented that are structured to show the level of confidence that the hypothesis is not rejected. Although they can never show that a model is right (model verification), these hypothesis tests can be used to show that a specific application of the model is not wrong. In this sense, model validation is a process of adding strength to our belief in the predictiveness of a model by repeatedly showing that it is not blatantly wrong in specific applications.

When a statistical test of observed versus predicted values fails to achieve the desired level of confidence, the problem may be with the observed data (usually the result of measurement error) or the model code (usually the specification of one or more key parameters). Recent developments in the statistical field of measurement error correction (*3*) have provided a tool for reducing the apparent effects of measurement error in the regression model.

In a single application of this measurement error correction procedure, this report has shown that hypothesis tests performed after measurement error correction can reverse the conclusion from rejection to acceptance of the statistical hypothesis, thus further validating the model and increasing the confidence that the model is not wrong. It is important to note that the measurement error correction procedure does not adjust any specific observation or drop any observation from the dataset. It uses the method of standard additions to adjust the slope and intercept of the regression between observed and predicted values.

Multiple regression models and related multiequation structural equation (pathway) models may require more sophisticated approaches. The study of measurement error effects using latent variable methods (*20*) is time consuming and labor intensive, requiring computer tests of several hours to days in length, using standard statistical packages such as SAS PROC CALIS (*18*). Unfortunately, intrinsically nonlinear models cannot be handled with existing packages.

There is also a need to evaluate and rank different model specification tests for empirical models when predictor variables are error-prone. Some recently developed methods for comparing different structural equation model specifications use residual curvilinearity (*21*). The effects of design matrix measurement errors on specification tests using residuals or studentized residuals from not-so-large samples is unknown. Cross-validation and bootstrap methods ought to be useful but may also need adjustments for measurement error effects.

REFERENCES AND NOTES

1. Bowers TS, Beck BD, Karam HS. Assessing the relationship between environmental lead concentrations and adult blood lead. Risk Anal 14:183–189 (1994).
2. U.S. EPA. Recommendations of the Technical Review Workgroup for Lead for an Interim Approach to Assessing Risks Associated with Adult Exposures to Lead in Soil. Washington:U.S. Environmental Protection Agency, 1996.
3. Carroll RJ, Ruppert D, Stefanski LA. Measurement Error in Nonlinear Models. London:Chapman & Hall, 1995.
4. U.S. EPA. Validation Strategy for the Integrated Exposure, Uptake, and Biokinetic Model for Lead in Children. EPA 540/R-94-039. Washington, DC:

U.S. Environmental Protection Agency, 1994.

5. Rabinowitz MB, Wetherill GW, Kopple JD. Lead metabolism in the normal human: stable isotope studies. Science (Washington) 182:725–727 (1973).
6. Rabinowitz MB, Wetherill GW, Kopple JD. Kinetic analysis of lead metabolism in healthy humans. J Clin Invest 58:260–270 (1976).
7. Rabinowitz M, Wetherill G, Kopple J. Delayed appearance of tracer lead in facial hair. Arch Environ Health 31:220–223 (1976).
8. Rabinowitz MB, Wetherill GW, Kopple JD. Magnitude of lead intake from respiration by normal man. J Lab Clin Med 90: 238–248 (1977).
9. Carroll RJ, Galindo CD. Measurement error, biases, and the validation of complex models for blood lead levels in children. Environ Health Perspect 106(Suppl 6):1535–1539 (1998).
10. U.S. EPA. Seasonal Rhythms of Blood-Lead Levels: Boston, 1979–1983. EPA 747/R-94-003. Washington:U.S. Environmental Protection Agency, 1995.
11. Hogan K, Marcus A, Smith R, White P. Integrated exposure uptake biokinetic model for lead in children: empirical comparisons with epidemiologic data. Environ Health Perspect 106(Suppl 6):1557–1567 (1998).
12. Mushak P. Uses and limits of empirical data in measuring and modeling human lead exposure. Environ Health Perspect 106 (Suppl 6):1467–1484 (1998).
13. Centers for Disease Control. Preventing Lead Poisoning in Young Children: A Statement by the Centers for Disease Control—October 1991. Atlanta:U.S. Department of Health and Human Services, 1991.
14. Kleinbaum DG, Kupper LL, Morgenstern H. Epidemiologic Research: Principles and Quantitative Methods. Belmont, CA:Lifetime Learning Publications, 1982.
15. Bishop YMM, Fienberg SE, Holland PW. Discrete Multivariate Analysis. Cambridge, MA:MIT Press, 1975.
16. Fuller WA. Measurement Error Models. New York:John Wiley & Sons, 1987.
17. Carroll RJ, Kuchenhoff H, Lombard F, Stefanski LA. Asymptotics for the SIMEX estimator in nonlinear measurement error models. J Am Stat Assoc 91:242–250 (1996).
18. SAS Institute, Inc. SAS/STAT User's Guide, Version 6, Fourth Edition, Vol 2. Cary, NC: SAS Institute Inc., 1990.
19. Marcus AH, Elias RW. Characterizing sources of variability in the US EPA IEUBK lead model by Monte Carlo simulation. Toxicologist 15:13 (1995).
20. Bollen KA. Structural Equations with Latent Variables. New York:John Wiley, 1989.
21. Jiang Q, Succop PA. A study of the specification of the log-log and log-additive models for the relationship between blood lead and environmental lead. J Agric Biol Environ Stat 1: 426–434 (1996).